УДК 519.24 : 658.15 : 628.148

Е.Е.ДАШЕВСКАЯ
ТПО «Харьковкоммунпромвод»

ПРИМЕНЕНИЕ МЕТОДОВ СТАТИСТИКИ И КЛАСТЕРИЗАЦИИ ТЕХНОЛОГИИ DATA MINING ПРИ ПОСТРОЕНИИ АСУ ПОТОКОРАСПРЕДЕЛЕНИЕМ В ИС

Анализируются методы технологии Data Mining и их применение для человеко-машинной оптимизации процедуры потокораспределением в ИС.

Для выработки оптимального и максимально эффективного решения в той или иной производственной ситуации в системе управления потокораспределением в инженерных сетях (ИС) огромное значение имеет интеграция накопленных производственных данных с современными аналитическими технологиями. Проблема выявления скрытых в массивах данных закономерностей стала актуальной, прежде всего потому, что данные, необходимые для принятия решения доступны, однако их так много и они столь сложно организованны, что принятие рационального решения затруднено. Выбор правильного решения зависит от представления данных в виде, четко отражающем производственные процессы, а также от построения модели, при помощи которой можно прогнозировать процессы, критичные для ИС.

Длительное время основным инструментом анализа данных служили традиционная математическая статистика и средства оперативной аналитической обработки данных. Следует отметить, что статистические методы используются для проверки заранее сформулированных гипотез. И именно формулировка гипотезы является самой сложной задачей при реализации производственного анализа для последующего принятия решения, поскольку не все закономерности в данных очевидны. В свою очередь извлечению соответствующих взаимосвязей экспертами- людьми препятствует субъективный фактор, что приводит к потребности в подходящих инструментах формализации знаний. В настоящее время таким инструментом признана технология Data Mining (добыча данных) в базах данных, которая на основании имеющегося набора данных способна находить взаимозависимости самостоятельно и строить гипотезы об их характере [1].

Сегодня сложные условия эксплуатации ИС не оставляют сомнения в необходимости исследовать накопленные данные предприятиями, используя передовые аналитические методы. В статье анализируются два метода исследования данных – суммарной статистики и кластеризации, а также аргументируется применение методов для построения прогнозирующей модели и интегрирование с набором про-

грамм поддержки принятия решения при автоматизации реальных производственных процессов в системах управления потокораспределением в ИС.

Задачи компьютерных систем поддержки принятия решений заключаются не только в фиксации опыта эксперта для выработки оптимального в данной ситуации решения, а и в анализе исторических данных. Под понятием исторические данные подразумевается комплекс показателей, описывающих поведение изучаемого объекта, принятых в прошлом решений, их результаты и т.д. Для того, чтобы применение этих систем на практике оказалось оправданным, необходимо репрезентативное множество этих данных – иначе принятые на их основе решения будут неосновательными. Исследование начинается с некоторой предварительной модели и ставится задача ее совершенствования для проверки найденных решений на корректность, значимость и допустимость [2].

Проанализируем выбранные методы в среде программы PolyAnalyst 4.6, в которой представлены и другие средства и методы технологии добычи данных.

Данные, представленные для анализа отражают факт аварийного повреждения на потокораспределительных сетях в течение года и представлены в табличной форме Excel. Количество записей является оптимальным в соответствии с требованиями программы PolyAnalyst 4.6 для наиболее эффективного применения алгоритмов исследования.

База данных содержит 4492 строки, и 9 параметров: место повреждения, местность повреждения, утечка ($м^3$/ч), диаметр, грунт, материал, вид, раскопка, количество дней ликвидации.

Рассмотрим один из часто применяемых методов анализа данных – статистический – Summary Statistics (SS). В результате выполнения SS – модуля был получен отчет для автоматически созданной таблицы WORLD, включающий два текстовых сообщения – описание числовых и целочисленных атрибутов (табл.1), характеристики категориальных атрибутов, а также диаграммы частотного распределения всех значений набора данных.

Таблица 1 – Числовые и целочисленные атрибуты таблицы WORLD

Числовые и целочисленные атрибуты	Значение	Среднее	Стд. откл.	Мин	Макс	Расстояние	Медиана
"Утеч $м^3$/ч"	3935	18,05	41,51	1	1000	999	5
Диаметр	4491	225,8	167,3	15	1500	1485	150
"Кол-во дней ликвидации"	4492	4,28	13,19	1	295	294	2

Числовые и целочисленные атрибуты описываются средним значением, стандартным отклонением, расстоянием, минимальным и максимальным значениями, медианой, модой (табл.1). Затем перечисляются все категории и атрибуты базы данных, сопровождающиеся числом значений, модой категорий, и списком распределения значений в пределах атрибута.

Краткая характеристика набора данных. Наибольшее количество аварийных ситуаций происходят на магистралях (79,39%), местность повреждения характеризуется как зеленая полоса (75,35%), грунт – глина (95,2%), для вида повреждения определено значение-свищ в 58,84% случаев, материал трубопровода – сталь (58,93%), 98,42% работ проводились с раскопкой. Среднее значение объема утечки в 1 ч составляет 18,05 $м^3$, срок ликвидации повреждения – 4 дня.

Этот метод необходимо использовать в начале анализа любого набора данных для очевидного и наглядного представления существующих взаимозависимостей.

В результате статистического анализа был получен краткий обзор данных, сравнение характеристик и распределений выбранного набора, сравнение частот категориальных атрибутов. Следует отметить тот факт, что этот метод не требует предварительной обработки данных и работает с любым типом данных и по степени наглядности сравниваемых атрибутов имеет приоритетное значение. Метод может использоваться для сравнения частот атрибутов между двумя наборами данных.

Выделим в базе данных компактные подгруппы записей (кластеры), проявляющие похожие свойства. Для этой цели в программе PolyAnalyst 4.6 есть специальный вычислительный модуль N – мерный кластеризатор Find Clusters (FC). Ниже приведен отчет о процессе кластеризации.

Из полученного отчета видно, что механизм поиска кластеров, учитывая все параметры таблицы, выделил три группы (табл.2), принадлежность к которым определяется двумя атрибутами: диаметром трубопровода и видом повреждения. Вероятность того, что кластеры обнаружены случайно, равна $7{,}3 \cdot 10^{-60}$, общее количество точек, входящих в кластеры 1469 записей из 4492. Первый кластер FC_World_1 содержит данные об аварийных повреждениях на стыках соединения, переломах, свищах, прочих и диаметром от 15 мм до 400 мм (1221 запись). Второй кластер FC_World_2 самый незначительный – 85 значений, касающиеся характера повреждения свища и диаметров более 650 мм. Третий кластер FC_World_3 – 163 записи, характер повреждения свища на диаметре трубопровода 500 мм.

Таблица 2 – Табличный отчет процесса кластеризации

Вид/Диам	(-,57.5)	(57.5,88)	(88, 125)	(125, 175)	(175,225)	(225, 275)	(275, 325)	(325, 375)	(375, 450)	(450, 550)	(550, 650)	(650,-)
Свищ points cluster	178 1	34 -	740 -	520 -	255 -	31 -	356 -	1 -	189 -	163 3	91 -	85 2
Перелом points cluster	7 -	4 -	190 1	164 1	99 1	20 1	41 -	12 1	2 -	2 -	2 -	0 -
Стык со points cluster	4 -	5 -	188 -	197 -	202 1	33 1	298 1	8 -	81 -	54 -	28 -	8 -
Прочие points cluster	22 1	3 1	8 -	0 -	2 -	0 -	1 -	0 -	1 -	1 -	0 -	0 -
Трещина points cluster	4 -	3 -	44 -	41 -	19 -	2 -	35 -	2 -	4 -	1 -	4 -	2 -

Таким образом, на основании полученных результатов можно сделать вывод, что наибольшее количество аварийных повреждений наблюдается на трубопроводах диаметром до 400 мм и, учитывая характер повреждения, они делятся на три крупные группы.

Оценка метода. Кластеризация находит подобные и аномальные области в наборе данных. Данный алгоритм не требует предварительной обработки данных и указания целевого атрибута, и в результате получается наглядная и конструктивная задача определения стратегии дальнейшего анализа данных.

Анализ рассмотренных методов исследования данных позволил сделать заключение о том, что их применение дает возможность максимально полно и точно раскрыть потенциал данных, быстро получить действенную информацию для эффективного управления и принятия решения. Сопоставляя характер анализируемых данных с полученными результатами можно утверждать, что рассмотренные методы технологии добычи данных являются неприхотливыми в плане состава анализируемых показателей, то есть достаточно слабый уровень организации первичных данных позволил получить достаточно конструктивные результаты.

Выбор необходимого метода исследования набора данных, на основании которого строится модель принятия решения, определяется содержательными особенностями задачи.

К перспективам исследований следует отнести работу по правильному интегрированию построенных прогнозирующих моделей в реальные производственные процессы для их оценки и адаптации к

другим задачам принятия решения.

1.Дюк В.А. Data Mining – интеллектуальный анализ данных. – СПб.: Питер, 2001. – 368 с.

2.Боровиков В. STATISTICA. Искусство анализа данных на компьютере. Для профессионалов. – 2-е изд. – СПб.: Питер, 2003. – 688 с.

Получено 04.03.2004

УДК 628.353

А.М.ТУГАЙ, д-р техн. наук, Я.А.ТУГАЙ, канд. техн. наук
Київський національний університет будівництва і архітектури

ВИЗНАЧЕННЯ ДОДАТКОВИХ НАПОРІВ ДЛЯ ЗБЕРЕЖЕННЯ ПОСТІЙНОЇ ПРОДУКТИВНОСТІ ВОДОЗАБІРНИХ СВЕРДЛОВИН В УМОВАХ ФІЛЬТРАЦІЙНИХ ДЕФОРМАЦІЙ ФІЛЬТРА І ПРИФІЛЬТРОВОЇ ЗОНИ

Розкривається механізм втрат напору при фільтраційних деформаціях фільтра і прифільтрової зони водозабірних свердловин та дається розрахунок визначення цих втрат при забезпеченні постійної продуктивності свердловин.

Практика експлуатації водозабірних свердловин (трубчастих колодязів) засвідчує зменшення їх продуктивності з часом в порівнянні з початковим. Це явище пов'язане з багатьма причинами, описаними в літературі, зокрема [1-4]. Однією із недостатньо досліджених причин цього явища є фільтраційні деформації фільтрів та прифільтрових зон свердловин, що ведуть до зміни щілинності (пористості) самих фільтрів і пористості породи безпосередньо біля свердловини. Основними видами фільтраційних деформацій, що ведуть до зменшення продуктивності свердловин є кольматаж фільтрів і прифільтрових зон свердловин. При цьому розрізняють механічний, хімічний і біологічний кольматаж. Фізико-хімічна природа кольматажу в прифільтровій зоні досить складна, оскільки тут відбуваються різні процеси, сутність яких викладено в спеціальній літературі, наприклад [1, 4, 5]. При цьому встановлено, що процеси кольматажу по своїй природі є однією із форм масообміну при фільтрації рідини в пористому середовищі, а тому можуть бути вивчені на основі теорії фільтрації, масопереносу і масообміну. Разом з тим, методи визначення кольматажу і оцінки його впливу на продуктивність водозабірних свердловин основані на занадто спрощених моделях і недостатньо повно відображають взаємопов'язані процеси фільтрації, масопереносу, масообміну і кінетики реакцій (при хімічному кольматажу). Тому гідродинамічна модель фільтрації в умовах наявності деформацій в прифільтровій зоні повинна складатися з двох взаємопов'язаних блоків: гідродинамічного (фільт-